Understanding the 2020 Census Disclosure Avoidance System:

Differential Privacy 101

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

May 4, 2021

Shape your future
START HERE >

United States Census 2020

Webinar Series:

Understanding the 2020 Census Disclosure Avoidance System

All webinars start at **1:00 pm EDT**

No pre-registration necessary. We will archive recordings to the website.

*Search *"Disclosure Updates"* at www.census.gov

Or link: https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html

Day	Date	Title
T	May 4	Differential Privacy 101
F	May 7	The Census Bureau's Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census
Th	May 13	Differential Privacy 201 and the TopDown Algorithm
F	May 14	Highlights of the April 2021 Detailed Summary Metrics
F	May 21	Analysis of April 2021 Demonstration Data for Redistricting and Voting Rights Act Use Cases

Acknowledgements

This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, including: John Abowd, Tammy Adams, Robert Ashmead, Craig Corl, Ryan Cummings, Jason Devine, John Fattaleh, Simson Garfinkel, Nathan Goldschlag, Michael Hawes, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Kyle Irimata, Dan Kifer, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Claudia Molinar, Brett Moran, Ned Porter, Sarah Powazek, Vikram Rao, Chris Rivers, Anne Ross, Ian Schmutte, William Sexton, Rob Sienkiewicz, Matthew Spence, Tori Velkoff, Lars Vilhuber, Bei Wang, Tommy Wright, Bill Yates, and Pavel Zhurlev.

For more information and technical details relating to the issues discussed in these slides, please contact the author at michael.b.hawes@census.gov.

Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.

The statistics included in this newsletter have been cleared for public dissemination by the Census Bureau's Disclosure Review Board (CBDRB-FY20-DSEP-001, CBDRB-FY20-281, and CBDRB-FY20-101).

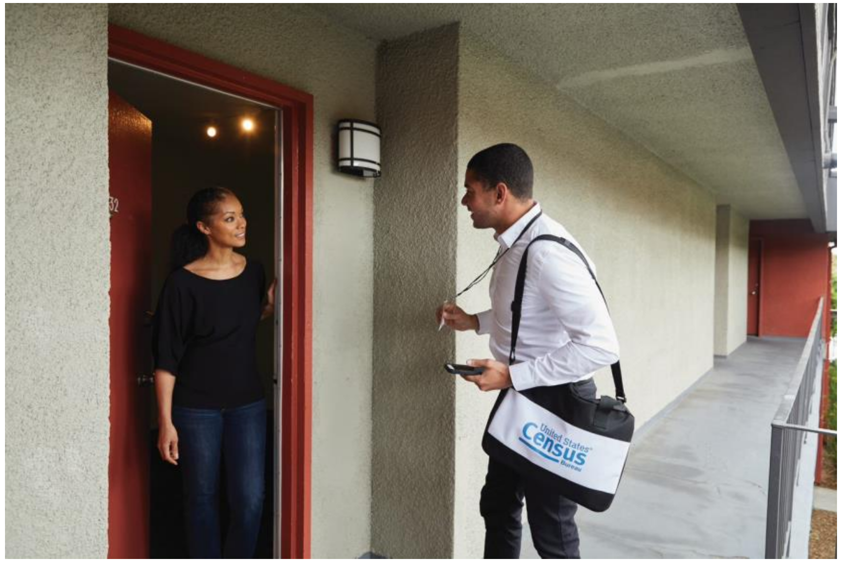
Shape your future START HERE >

United States Census 2020

Our Commitment to Privacy and Confidentiality

Data stewardship is central to the Census Bureau's mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



Shape your future START HERE >

United States® Census 2020

Upholding our Promise: Today and Tomorrow

We cannot merely consider privacy threats that exist today.

We must ensure that our disclosure avoidance methods are also sufficient to protect against the threats of tomorrow!

The Census Bureau's Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.

Stopped publishing small area data	Whole-table suppression	Data swapping	Differential Privacy
1930	1970	1990	2020

The Privacy Challenge

Every time you release any statistic calculated from a confidential data source you "leak" a small amount of private information.

If you release too many statistics, too accurately, you will eventually reveal the entire underlying confidential data source.

Dinur, Irit and Kobbi Nissim (2003) "Revealing Information while Preserving Privacy" PODS, June 9-12, 2003, San Diego, CA

The Growing Privacy Threat

More Data and Faster Computers!

In today's digital age, there has been a proliferation of databases that could potentially be used to attempt to undermine the privacy protections of our statistical data products.

Similarly, today's computers are able to perform complex, large-scale calculations with increasing ease.

These parallel trends represent new threats to our ability to safeguard respondents' data.

Reconstruction

The recreation of individual-level data from tabular or aggregate data.

If you release enough tables or statistics, eventually there will be a unique solution for what the underlying individual-level data were.

Computer algorithms can do this very easily.

	4						2	
			7					4
1		7	8				5	
			9			3		8
5								
			6		8			
3						4		5
	8	5				1		9
		9		7	1			

Reconstruction: An Example

	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

Reconstruction: An Example

	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married
36	Female	Black	Married
8	Female	Black	Single
18	Male	White	Single
24	Female	White	Single

This table can be expressed by 164 equations.
Solving those equations takes 0.2 seconds on a 2013 MacBook Pro.

Re-identification

Linking public data to external data sources to re-identify specific individuals within the data.

Name	Age	Sex
Jane Smith	66	Female
Joe Public	84	Male
John Citizen	30	Male

External Data

+

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married

Confidential Data

Reconstructing the 2010 Census

- The 2010 Census collected information on the age, sex, race, ethnicity, and relationship (to householder) status for ~309 Million individuals. (1.9 Billion confidential data points)
- The 2010 Census data products released over 150 billion statistics
- We conducted an internal experiment to see if we could reconstruct and re-identify the 2010 Census records.

Shape your future START HERE >

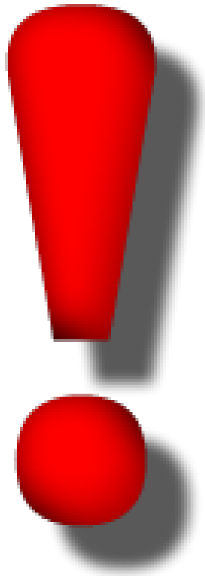
United States Census 2020

Reconstructing the 2010 Census: What Did We Find?

1. On the 309 million reconstructed records, census block and voting age (18+) were correctly reconstructed for all individuals in all 6,207,027 inhabited blocks.
2. Block, sex, age (in years), race (OMB 63 categories), and ethnicity were reconstructed:
    1. Exactly for 46% of the population (142 million individuals)
    2. Within +/- one year for 71% of the population (219 million individuals)
3. Block, sex, and age were then linked to commercial data, which provided presumed re-identification of 45% of the population (138 million individuals).
4. Name, block, sex, age, race, ethnicity were then compared to the confidential data, which yielded confirmed re-identifications for 38% of the presumed re-identifications (52 million individuals).
5. For the confirmed re-identifications, race and ethnicity are learned correctly, though the attacker may still have uncertainty.

The Census Bureau's Decision

- Advances in computing power and the availability of external data sources make database reconstruction and re-identification increasingly likely.
- The Census Bureau recognized that its traditional disclosure avoidance methods are increasingly insufficient to counter these risks.
- To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections.



Disclosure Avoidance

Disclosure avoidance methods seek to make reconstruction and re-identification more difficult, by:

- Reducing precision
- Removing vulnerable records, or
- Adding uncertainty

Commonly used (legacy) methods include:

- Complementary suppression
- Rounding
- Top/Bottom coding of extreme values
- Sampling
- Record swapping
- Noise injection

Problem #1 – Impact on Data

All statistical techniques to protect privacy impose a tradeoff between the degree of privacy protection and the resulting accuracy of the data.

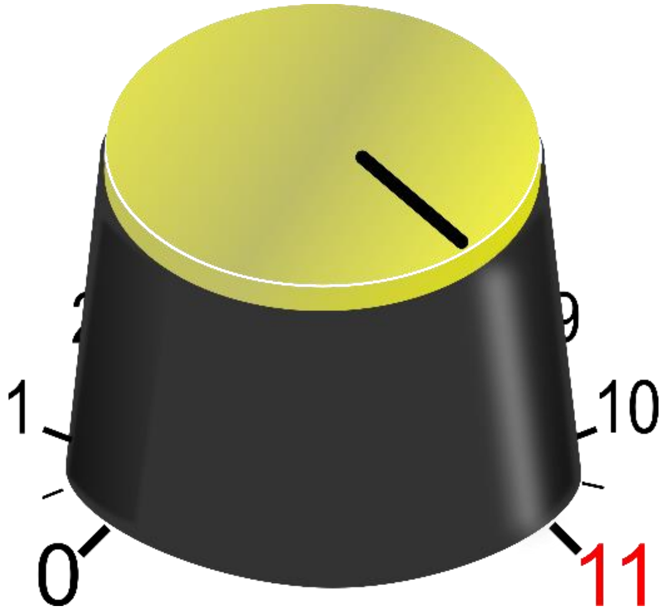
Swap rates, noise injection parameters, cell suppression thresholds, etc. determine this tradeoff.

Problem #2 – How much is enough?

Legacy disclosure avoidance methods provide little ability to quantify privacy protections.

When faced with rising disclosure risk, disclosure avoidance practitioners adjust their implementation parameters.

BUT, this is largely a scattershot solution that over-protects some data, while often under-protecting the most vulnerable records.



Differential Privacy

DP is not a disclosure avoidance "method" as much as it is a framework for <u>defining</u> and then <u>quantifying</u> privacy protection.

- Every individual that is reflected in a particular statistic contributes towards that statistic's value.
- Every statistic that you publish "leaks" a small amount of private information.
- DP as a framework allows you to assess each individual's contribution to the statistic, and to measure (and thus, limit) how much information about them will leak.

Differential Privacy

When combined with noise injection, DP allows you to precisely control the amount of private information leakage in your published statistics.

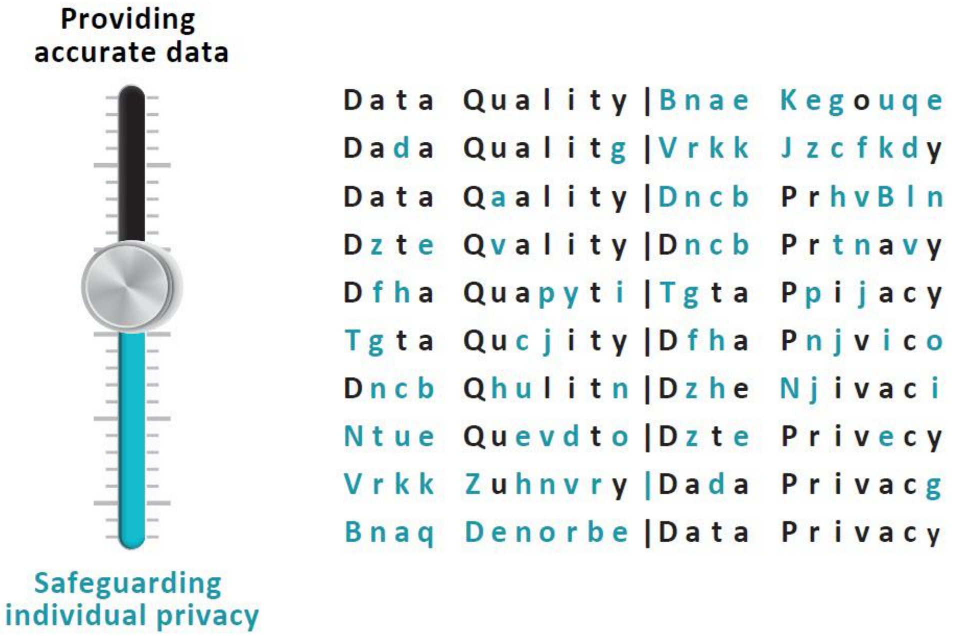
- **Infinitely tunable – parameter "dials" can be set anywhere from perfect privacy to perfect accuracy.**
- **Privacy guarantee is mathematically provable and future-proof.**
- **The precise calibration of statistical noise enables optimal data accuracy for any given level of privacy protection.***

***Absent post-processing requirements, which can introduce error independent of that needed to protect privacy.**

Privacy vs. Accuracy

The only way to absolutely eliminate all risk of re-identification would be to never release any usable data.

Differential privacy allows you to quantify a precise level of "acceptable risk," and to precisely calibrate where on the privacy/accuracy spectrum the resulting data will be.



Establishing a Privacy-loss Budget

This measure is called the "Privacy-loss Budget" (PLB) or "Epsilon."

- ε=0 (perfect privacy) would result in completely useless data
- ε=∞ (perfect accuracy) would result in releasing the data in fully identifiable form

ε

Epsilon

Comparing Methods

<u>Data Accuracy</u>

Differentially private disclosure avoidance methods are not inherently better or worse than traditional methods.

Both can have varying degrees of impact on data quality depending on the parameters selected and the methods' implementation.

<u>Privacy</u>

Differentially private methods are substantially better than traditional methods for protecting privacy, insofar as they actually allow for measurement of the privacy risk.

Implications for the 2020 Census

The modernization of our privacy protections using a differential privacy framework does not change the constitutional mandate to apportion the House of Representatives according to the actual enumeration.

As in 2000 and 2010, the Census Bureau will apply privacy protections to the P.L. 94-171 redistricting data.

Shape your future START HERE >

United States Census 2020

Privacy-loss Budget Allocation

The Census Bureau's Data Stewardship Executive Policy Committee (DSEP) will be making decisions about the PLB for the 2020 Census. This includes allocation across different 2020 Census data products, including:

- P.L. 94-171 Redistricting data
- Demographic and Housing Characteristics files (DHC)
- Detailed Demographic and Housing Characteristics files (D-DHC)
- …and other uses of Decennial Census data.

DSEP will also be deciding how to allocate the PLB across the different sets of tabulations *within* each data product (by geographic level and by data element).

Recent Activity: DAS Tuning for the Redistricting Data

P.L. 94-171 Tuning & Privacy-Accuracy Trade-off Experiments

- In December through March, the DAS Team conducted over 600 full-scale TDA runs with the complete P.L. 94-171 data product schema.
- Goal: Evaluating resulting accuracy of varying parameters for:
  - Overall setting of PLB
  - Query strategy
  - Allocation of PLB across geographic levels
  - Allocation of PLB across queries
- Worked with subject matter experts in Demographic and Decennial Directorates to evaluate accuracy of experimental runs to inform parameter setting.

Demonstration Data

- Since October 2019, the Census Bureau has been periodically releasing demonstration data products (using 2010 Census data) for data user evaluation.
- The first four of these sets of demonstration data (October 2019, May 2020, September 2020, November 2020) used a conservative global PLB set by DSEP for the October 2019 Demonstration Product, in order to evaluate algorithmic improvements.
- ***The 2020 Census Data Products will not be held to this fixed PLB.***

- On April 28, we released another set of Privacy-Protected Microdata Files (PPMFs) and Detailed Summary Metrics using a different global PLB (ε=12.2) that more closely approximates the level of PLB that the DSEP will be considering for the 2020 Census redistricting data files.
- In September, we plan to release a final set of PPMFs using the actual production code and settings that will be used for the 2020 Census redistricting data files.

How to Submit Feedback

The changes in the April 2021 PPMFs data set reflect the cumulative feedback received from the data user community throughout the development process. We look forward to feedback from data users on this new demonstration product. Your input will inform the Census Bureau's June 2021 final decision on the PLB and on the 2020 Census redistricting data parameters. **The deadline to submit feedback is May 28, 2021.**

**** Please send comments to 2020DAS@census.gov with the subject line "April 2021 Demonstration Data."**

Particularly useful feedback would describe:

•**Fitness-for-use:** Based on your analysis, would the data needed for your applications (redistricting, Voting Rights Act analysis, estimates, projections, funding data sets, etc.) be satisfactory?

- How did you come to that conclusion?
- If your analysis found the data to be unsatisfactory, how incrementally would accuracy need to change to improve the use of the data for your required or programmatic use case(s)?
- Have you identified any improbable results in the data that would be helpful for us to understand?"

•**Privacy:** Do the proposed products present any confidentiality concerns that we should address in the DAS?
•**Improvements:** Are there improvements you've identified that you want to make sure we retain in the final design? Be specific about the geography and error metric for the proposed improvement.

Stay Informed:

Subscribe to the 2020 Census Data Products Newsletters

*Search “Disclosure Avoidance” at www.census.gov

Stay Informed:

Visit Our Website

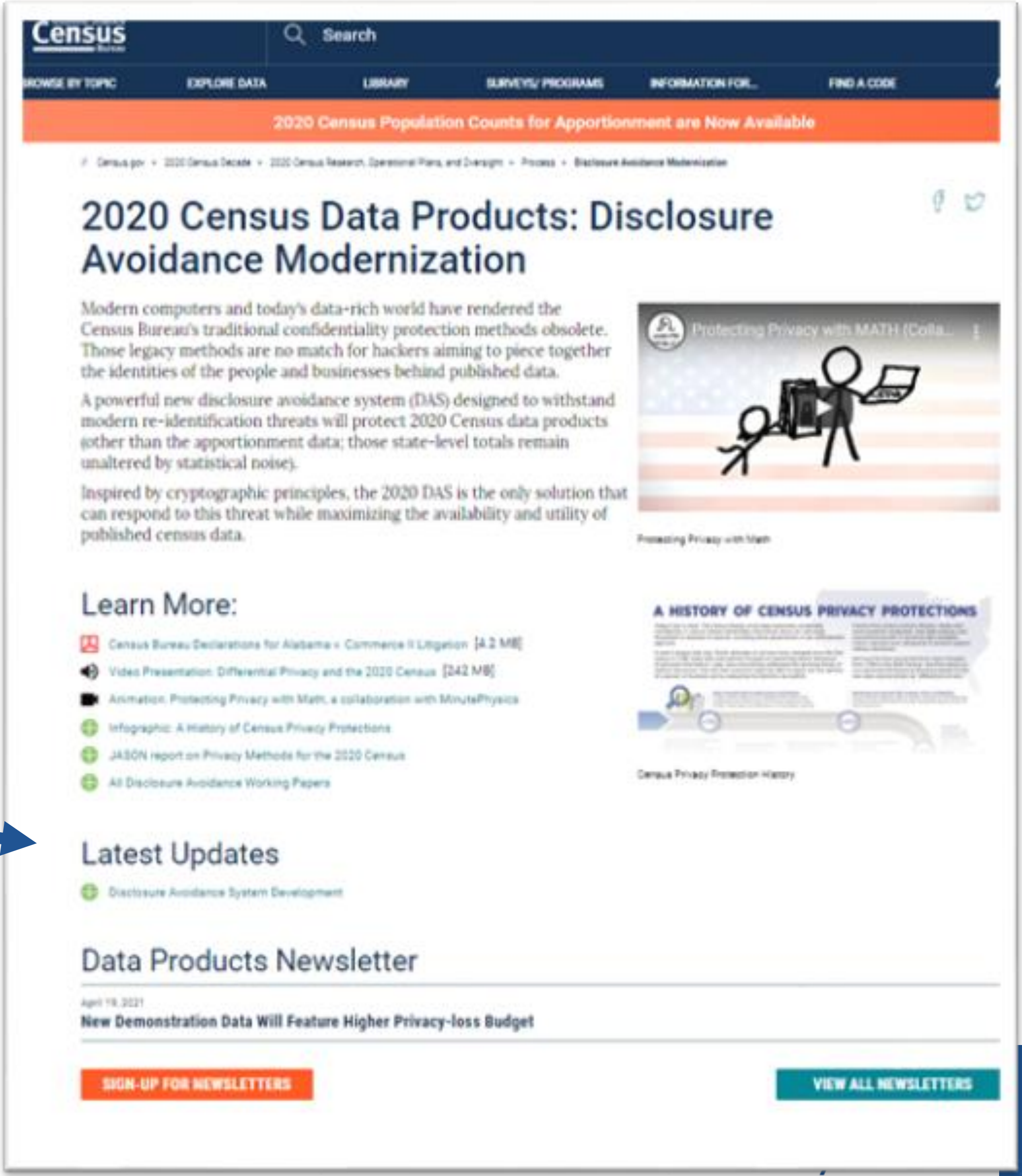
*Search “Disclosure Avoidance” at www.census.gov

Latest Updates

Disclosure Avoidance System Development

Dates and log-In information for webinar series

coming soon



START HERE >

Questions?

Shape
your future
START HERE >

United States®
Census
2020